Observational Studies () Submitted ; Published

Why are RCTs the Gold Standard? The Epistemological Difference Between Randomized Experiments and Observational Studies

Christopher Harshaw **c.harshaw@columbia.edu**
Department of Statistics
Columbia University
New York City, NY 10027, USA

Abstract

In response to Pearl, Aronow et al. (2025) argue that randomized experiments are special among causal inference methods due to their statistical properties. I believe that the key distinction between randomized experiments and observational studies is not statistical, but rather epistemological in nature. In this comment, I aim to articulate this epistemological distinction and argue that it ought to take a more central role in these discussions.

Keywords: Causal inference, Randomized experiments, Epistemology

1. Introduction

I thank the editors for the invitation to comment on the paper by Aronow et al. (2025). The authors should be congratulated for their clear contribution to the ongoing discussion regarding the distinction between causal inference methods. These discussions are especially important as causal inference methods are becoming more widely adopted in a variety of disciplines. As such, I am honored to be able to contribute to this vibrant and timely dialogue.

The authors are reacting to Judea Pearl, who asserts that statistical identification is what makes a causal inference method "work" and that from this perspective, "there is no need to put [randomized experiments] on a pedestal". Aronow et al. argue that randomized experiments ought to be considered special among causal inference methods because they guarantee precise statistical estimation under weaker conditions than what is required in an observational study. At the core of their argument is the assertion that precise statistical estimation—not merely statistical identification—should be a determining factor for whether a causal inference method "works". The reasoning goes that the analyst will, after all, face the task of analyzing actual data and not the inaccessible observational distribution.

I wholeheartedly agree with the conclusion of Aronow et al. (2025) that randomized experiments ought to be considered special among causal inference methods. However, it is my opinion that these statistical distinctions are of secondary concern. Instead, I believe that the most meaningful distinction between randomized experiments and observational studies is an epistemological one. In the remainder of my comment, I hope to clarify this epistemological distinction and argue that it ought to take a more central role in deciding whether a causal inference method "works".

That randomized experiments offer superior types of empirical evidence compared to observational studies seems to be a widely held belief among methodologists and applied researchers alike.[1] Randomized experiments are almost ubiquitously referred to as the "gold standard" among causal inference methods, to such an extent that this phrase has effectively become cliché. In response to critiques of causal inference methods in economics (Deaton, 2009; Heckman and Urzúa, 2010), Imbens (2010) writes that "randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top" due to their high degree of "internal validity". Freedman (2006) asserts that "experiments offer more reliable evidence than observational studies".

While I agree with the general conclusion, what I find to be missing from this general sort of argument is a clear articulation of why randomized experiments have this particular distinction. What exactly gives randomized experiments such high credibility and why do observational studies not have it? I claim that the answer lies in how we argue for the validity of the resulting conclusions.

2. The Epistemological Distinction

Causal inference methods are primarily deductive; that is, if each of the prerequisite assumptions are true and the method is carried out in the prescribed manner, then the statistical conclusions regarding the treatment effect will be valid.[2] Like all deductive methods, the validity of these conclusions relies on establishing the validity of the assumptions, which is referred to as the "weakest link" principle (Cartwright, 2007). Two very common assumptions in causal inference methods are *unconfoundedness* (i.e. outcomes and treatment are conditionally independent given confounding covariates) and *positivity* (i.e. each individual has a positive probability of receiving treatment). These two assumptions are concerned with the process by which treatment was allocated to or chosen by the participants in the study, which is commonly referred to as the *treatment assignment mechanism* in the causal inference literature. In order to ensure that the causal conclusions are valid, the analyst must convincingly argue that these two assumptions are themselves valid.

In a randomized experiment, these two assumptions are easily shown to be valid. In particular, the treatment assignment mechanism was designed and carried out by the experimenter so that its description and proper execution are enough to ensure that these two assumptions hold. To verify unconfoundedness, the experimenter recounts how they assigned treatment blindly and without regards for potential outcomes, so that independence between these two quantities must hold. To verify positivity, the experimenter needs only to recall how they assigned treatment by flipping coins which had a positive probability of coming up heads and similarly for tails.

Things are quite different in an observational study, where the verification of these statistical assumptions is a fundamentally different task. By the nature of an observational

1. Cartwright (2007) is a dissenting voice but her critiques apply broadly to all statistical methods. In particular, she does not engage with the actual statistical assumptions which demonstrate the advantages of a randomized experiment over an observational study.
2. Statistical conclusions about the treatment effect (e.g. confidence intervals, hypothesis tests) are narrower in scope than exact conclusions because they will always reflect the fundamental limitation that estimands can be known only up to some amount of error. In the interest of clarity, I will refer to these types of statistical conclusions simply as "conclusions".

study, very little is known about the treatment assignment mechanism. For this reason, it will typically be impossible to argue directly for the validity of these assumptions by reference to some known material process, as is the case in a randomized experiment. How then is the analyst supposed to argue for the reasonable plausibility—let alone validity—of these statistical assumptions?

When this question arises, the standard methodological response is to say that the verification of these assumptions should be left to subject matter expertise. But what exactly does this mean? Ideally, the expert analyst will use previously collected evidence, introspection, and reasoning to craft a thought experiment which explains the existence of the treatment assignment mechanism and justifies the assumptions. This thought experiment may or may not be entirely convincing to the intended audience. If it is, the conclusions of the study are considered to be valid. If it is not, then subject matter expertise has failed to justify the perquisite assumptions and thus the conclusions are not considered valid. This is the reality of appealing to subject matter expertise in an observational study.

Drawing meaningful conclusions from an observational study relies on an expert analyst to construct a convincing *story* for why the treatment assignment mechanism ought to satisfy the prerequisite assumptions. On the other hand, the conclusion of a randomized experiment are valid by the experimenter's deliberate act of randomization. These justifications offer fundamentally different kinds of credibility. The latter is neither an analogy, nor a thought experiment, nor any other rhetorical device—it is something that actually happened. Put simply: it is always more credible to recount an actual chain of events than to convince people of the veracity of a fanciful story, even if expertly constructed.

Randomized experiments are indeed special among causal inference methods and this epistemological distinction is the strongest argument to that end. After all, the analyst will be faced with the task of convincing their peers of their empirical findings. In light of this epistemological difference, statistical concerns such as identification and uniform consistency—though certainly important—seem like minor details.

I do not mean to downplay the importance, relevance, and usefulness of observational studies and their findings. My opinion is in fact quite the opposite. I believe that we benefit greatly from rigorous observational studies. Certain social and scientific phenomenon do not permit control and are only accessible by means of passive observation. Moreover, it is often unethical or prohibitively costly to run a randomized experiment, in which cases observational studies are the only causal inference methods available. But this should not be taken to mean that observational studies are easier to successfully execute than randomized experiments. Providing sufficient evidence for the justification of the additional assumptions required in an observational study requires great efforts on the part of the analyst. My point is merely that we ought to recognize the epistemological difference as we use causal inference methods to test, corroborate, falsify, and otherwise contribute to scientific and social theories.

3. Design-Based Perspective

Absent from this discussion has been the canonical statistical framework itself: the *super-population framework*. This framework posits the assumption that subjects in the study are are independent and identical samples from a super-population (i.e. probability measure).

Although this may seem like an innocuous assumption to those with conventional statistical training, it can be the source of a great deal of confusion: which actual group of people in the world are being referenced by this mathematical abstraction of the super-population? How would we verify that subjects were indeed independent and identical representatives of this group?

The fact of the matter is that a super-population assumption is never supposed to be interpreted as being literally true. Instead, super-populations are once again stories that are carefully constructed by subject matter experts. The super-population framework is the predominant mode of statistical analysis used in practice and taught in the classroom. When analyzing randomized experiments within this framework, their key epistemological advantage is lost. This is likely the reason why the epistemological advantage has not been widely recognized.

In response to these issues, there has been a recent but steadily growing literature which is referred to as *design-based inference*. To the best of my knowledge, the design-based perspective was first used by Neyman (1923) and revived in the modern literature by Freedman (2008) with subsequent work from several scholars (see, e.g. Lin, 2013; Aronow and Samii, 2017; Harshaw et al., 2022). The essential feature of this body of work is that the super-population assumption is rejected and, instead, the subjects in the study are considered to be fixed. The causal effect to be investigated is defined only on the subjects in the experiment. Treatment assignment is the sole source of randomness within this framework and it serves as the primary basis for statistical inference. Although there are interesting statistical considerations that arise in a design-based setting, the primary benefit of this framework is epistemological. Rather than relying on a thought experiment in order to justify a super-population assumption, the design-based perspective is clearer in its aims. The design-based framework does not require stories, or at least less fanciful ones.

Some methodologists view the design-based framework as being too restrictive, unduly limiting the scope of the study. For example, they might point to the fact that causal conclusions reflect only the subjects in the study. However, I would argue in light of the concerns above that this restriction actually serves to strengthen the validity of the conclusions and clarifies exactly what sort of knowledge can be credibly ascertained from a randomized experiment. To this end, I would ask them: is the design-based conclusion actually narrower in scope if the audience does not believe in your story? It is rarely clear which group of people the super-population is supposed to capture. The only reason researchers do not feel more unsettled about this common confusion is an ignorance bred by familiarity.

Imbens (2010) highlights that causal inference methods attain their credibility through the research design rather than relying on particular modeling assumptions. This sentiment characterizes the general ethos of causal inference research. While matters of statistical estimation and identification are of great importance, they are not the sole determinants of credibility. Instead, it is the epistemological concerns—namely, the types of storytelling required to validate empirical results—that ultimately determine credibility. From this perspective, design-based inference takes the spirit of causal inference to its most extreme form.

Acknowledgments

I would like to thank P.M. Aronow, Lily Hu, Issa Kohler-Hausmann, Betsy Ogburn, Ben Recht, James M. Robins, Andrea Rotnitzky, and especially Fredrik Sävje for insightful discussions which have profoundly shaped my understanding of the epistemological foundations of causal inference and statistics.

References

P.M. Aronow and Cyrus Samii. Estimating average causal effects under general interference. *Annals of Applied Statistics*, 11(4):1912–1947, 2017.

P.M. Aronow, James M. Robins, Theo Saarinen, Fredrik Sävje, and Jasjeet Sekhon. Nonparametric identification is not enough, but randomized controlled trials are. *Observational Studies*, 2025.

Nancy Cartwright. Are RCTs the gold standard? *BioSocieties*, 2(1):11–20, 2007.

Angus Deaton. Instruments of development: Randomisation in the tropics, and the search for the elusive keys to economic development: Keynes lecture in economics. In *Proceedings of the British Academy, Volume 162, 2008 Lectures*. British Academy, 2009.

David A. Freedman. Statistical models for causation: what inferential leverage do they provide? *Evaluation Review*, 30(6):691–713, 2006.

David A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

Christopher Harshaw, Fredrik Sävje, and Yitan Wang. A design-based Riesz representation framework for randomized experiments. arXiv:2210.08698, 2022.

James J. Heckman and Sergio Urzúa. Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, 156(1):27–37, 2010.

Guido W. Imbens. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423, 2010.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1):295–318, 2013.

Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1923. This republication appeared in 1990.